ARMY RESEARCH LABORATORY

An Entropy-based Approach to Detecting Anomalies in Voice over Internet Protocol (VoIP) Traffic

by Gardner W. Thompson

ARL-TR-5124 **March 2010**

Approved for public release; distribution unlimited.

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer’s or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory

Aberdeen Proving Ground, MD 21005

ARL-TR-5124 **March 2010**

An Entropy-based Approach to Detecting Anomalies in Voice over Internet Protocol (VoIP) Traffic

Gardner W. Thompson
Computational and Information Sciences Directorate, ARL

Approved for public release; distribution unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED (From - To)**
March 2010	Final	5/19/2009 — 12/29/2009

4. TITLE AND SUBTITLE An Entropy-based Approach to Detecting Anomalies in Voice over Internet Protocol (VoIP) Traffic	**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Gardner W. Thompson	**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Laboratory ATTN: RDRL-CIN-D Aberdeen Proving Ground, MD 21005	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-TR-5124
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)** **11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Computer intrusion is a growing concern and field of investigation among government and private agencies. The main issue with most of the current Intrusion Detection Systems (IDSs) is that they are based on signature based observations, which means this class of detection system will only alert on attacks that the system is programmed to see. Entropy can be applied in various ways to examine data, but it is not a standalone IDS. It offers a theoretical, yet practical approach for the detection of abnormal patterns of behavior.

15. SUBJECT TERMS

Entropy, voice over internet, intrusion

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	UU	18	Gardner W. Thompson
Unclassified	Unclassified	Unclassified			**19b. TELEPHONE NUMBER** *(Include area code)* (410) 278-6509

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

1. Introduction

Computer intrusion is a growing concern and field of investigation among government and private agencies. The main issue with most of the current Intrusion Detection Systems (IDSs) is that they are based on signature based observations, which means this class of detection system will only alert on attacks that the system is programmed to see. This technical report investigates the use of entropy for detecting computer anomalies. Using entropy will allow us to detect strange occurrences within a given timeframe. One example of using entropy is to potentially detect data exfiltration using packet size distribution. The U.S. Army Research Laboratory (ARL) is investigating the ex-filtration of data using unbounded fields in Voice over Internet Protocol (VoIP) Session Initiation Protocol (SIP) packets. Entropy offers a theoretical approach for the detection of abnormalities in the protocol which could be indicative of malicious behavior.

2. Background

2.1 Entropy

Entropy has several different definitions (*1*). Shannon's definition of entropy is the most commonly used and the one used in this paper.

$$E = \sum_{i=1}^{n} p_i \log 2(p_i)$$

In the above formula, there are n events and the probability of the ith event is p_i. Note that the values of the events do not influence the value of the entropy only the probabilities are of concern. Changes in entropy will reflect a change in the set of probabilities representing the event space. Event spaces with different values but the same set of probabilities will be equivalent from the perspective of entropy. Entropy is a good way to detect suspicious behavior over a period of time. When strange activity has been detected during a time frame, we must use some type of anomaly detection tool to find the individual event.

2.2 VoIP

VoIP is an up and coming technology that gives both foreign and domestic enemies new ways to transmit hidden messages or infiltrate a network via VoIP technology (*2*). It functions by letting its users talk over the internet using phone to phone, computer to computer, or computer to phone communication devices.

Figure 1 demonstrates how VoIP operates by converting voice to a digital signal which travels over the internet. If the call is directed to a phone, the signal is translated into a regular phone signal when it reaches its destination. A broadband connection is also required to be able to use VoIP.

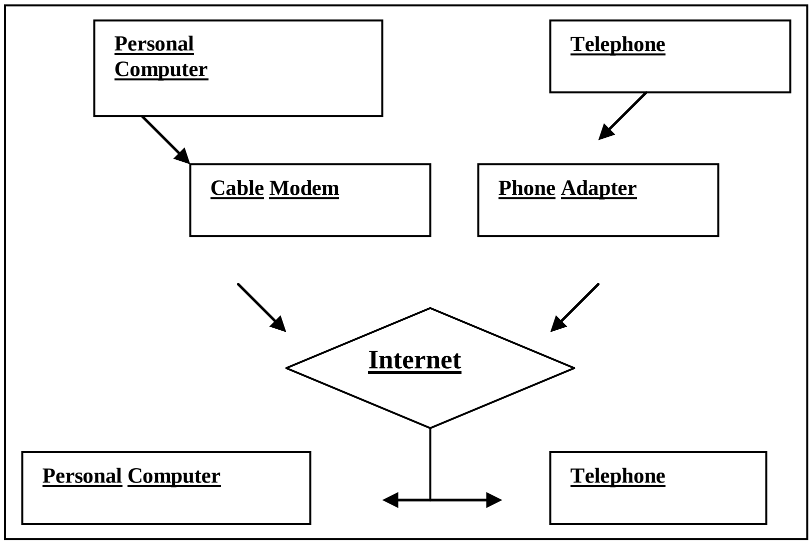

Figure 1. High level view of a typical VoIP architecture.

3. Method

For this research, we calculated the entropy of a series of SIP packets, focusing specifically on the distribution of packet sizes per SIP packet type. The packets we used for our research were REGISTER, INVITE, and CANCEL.

- REGISTER: Used by a user authentication (UA) to notify its current Internet Protocol (IP) address and the Uniform Resource Locator (URLs) for which it would like to receive calls.
- INVITE: Used to establish a media session between user agents.
- ACK: Confirms reliable message exchanges.
- CANCEL: Terminates a pending request.
- BYE: Terminates a session between two users in a conference.
- OPTIONS: Requests information about the capabilities of a caller, without setting up a call.

Details about the SIP, ACK, BYE, and OPTIONS packet types follow in the paragraphs below.

The metric we used for our analysis is packet size (*3*). This metric was the one that allowed us to distinguish between each packet type mentioned above. Other metrics that we considered using were IP addresses and port numbers. These, however, did not provide any information to identify the different types of SIP packets.

Entropy was used to investigate packet size; a packet is the basic information unit on a network. All communications are pieced into packets. The packets examined in this report are Register, Invite, and Cancel packets. Packet sizes change based on the amount of data an individual packet carries (i.e., more data will result in a bigger packet size). When examining the size of these packets we considered the type of packet it was. In the VoIP data we used for our analysis, the INVITE packet sizes range from 550 to 1076 bytes, the CANCEL packets range from 375–609 bytes, and REGISTER packets range from 302–680 bytes.

4. Simulations and Results

It is important to be able to identify significant changes in entropy. Simulations were designed so that significant entropy changes could be determined. Several simulations were done to analyze the affect that changing the number of observations and Unique Packet Sizes (UPS) has on the overall entropy of a data set. By observing the variation in entropy caused by random sampling, it is possible to determine the differences in entropy that are considered significant. This must be taken into consideration in order to minimize the false alarm rate.

The box plots in figure 2 show how the entropy changes as the number of UPS change and the number of observations stay the same. Each box plot represents 100 replications of the entropy for a set of 800 randomly chosen observations with the indicated number of categories. The range of the entropy through the entire graph is only .25, but when the details are examined closer it is easy to see the amount of change. The red line represents the median entropy and the blue outer box represents 50% of the data that is closest to the median or the inner quartiles. The whiskers represent the rest of the data in the upper and lower quartile, respectively. It is clear to see that when adding or subtracting 5 bytes to the UPS the entropy only changes gradually. For example, the median entropy for 185 UPS lies within the box plot that only contains 180 UPS. On the other hand, when 180 UPS are compared to 190 UPS you can see that there is an obvious change in entropy because the median of 190 UPS is not within the box plot of 180. So when the number of UPS is changed by 10 and the number of observations stays the same the entropy has a clear change. As well as using the inner quartile ranges as an indicator, a possible detection criterion could be based on the standard deviation of the entropy. This method could be used to detect statistical anomalies of a data set.

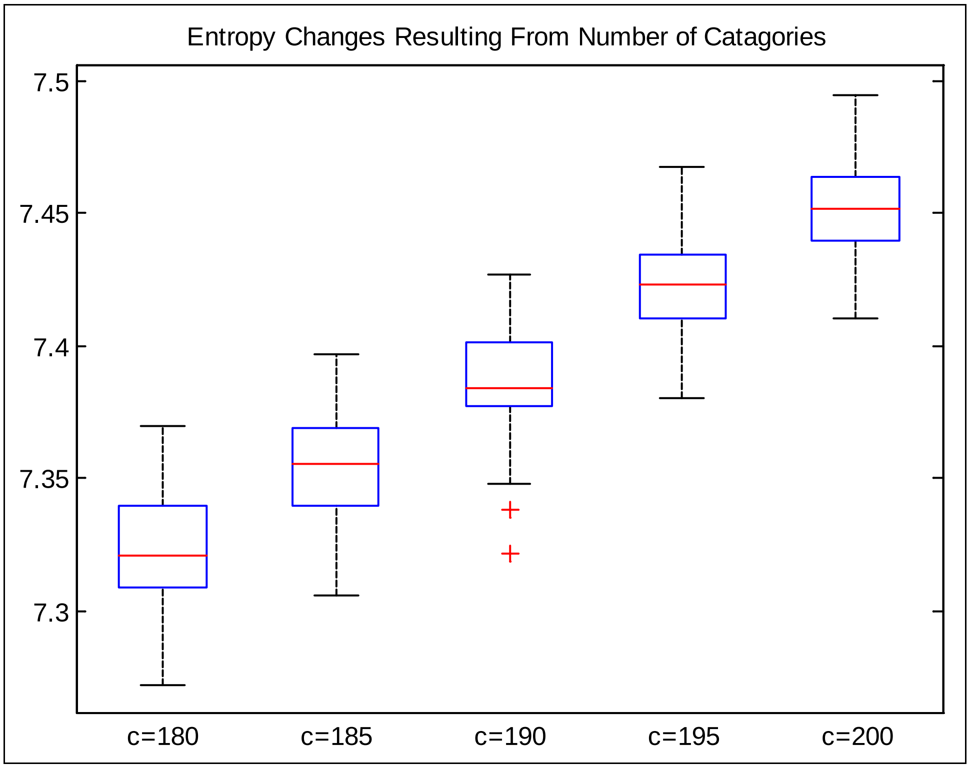

Figure 2. Entropy change for 800 observations based on different number of categories.

When examining the entropy through the change in the amount of observations and leaving the UPS the same (as shown in figure 3), the first thing you see is that as the number of observations increase the variation in the entropy decreases. This is expected because the UPS is staying the same and the number of observations for each entropy calculation is going up, so the only way for the variation to go is down.

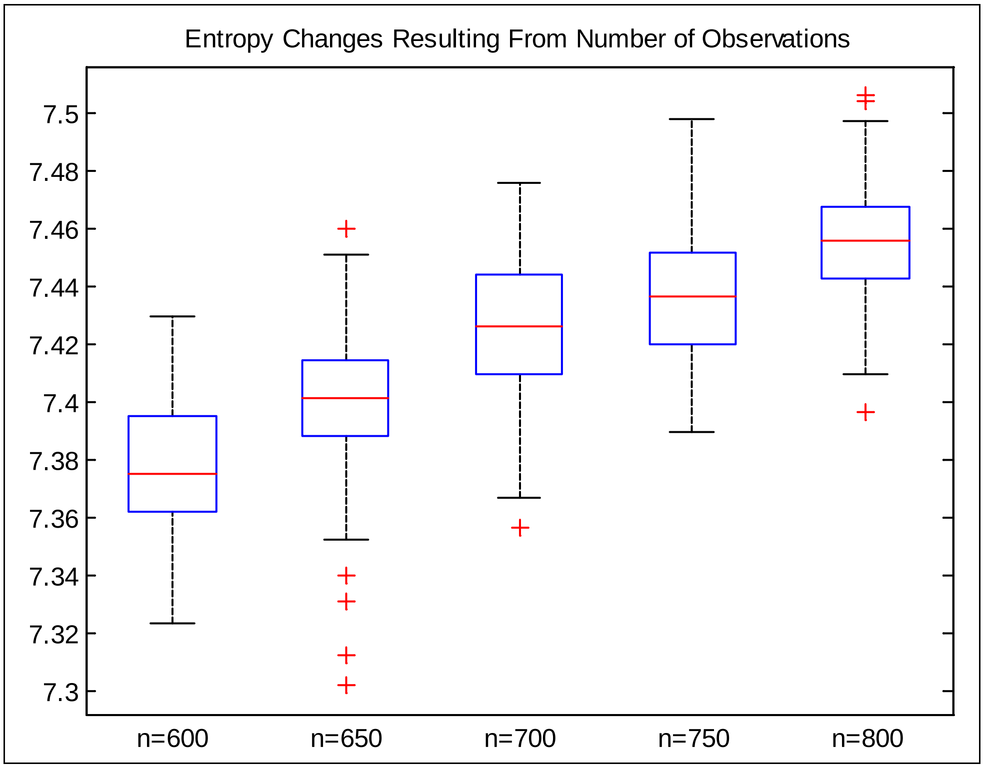

Figure 3. Changes in entropy as a function of sample size.

As the observations increase the entropy also increases. There is very little variation in the entropy when changing the number of observations. This means that a modest change in the number of observations does not have a large affect on the entropy. For example, when reviewing figure 3, the variation in the entropy median from 600 observations to 800 observations; or a 33% increase in observations, only has a 0.08 increase in entropy. This can be applied to other situations. For example, the entropy of a set with 10,000 observations could be compared to the entropy of a data set which includes 9,000 observations (which is only a 10% difference in observations); and it would be reasonable to make decisions based on the difference in entropy.

5. Data Description

For this research, we used a packet dump that consisted of 82 packets. We then categorized them based on SIP packet type (i.e., INVITE, REGISTER, CANCEL). Due to the small number of packets, we combined the packets together and bootstrapped them. To bootstrap is to choose randomly sampled points with replacement from the data set, and then analyzing them using the same method. With replacement, it means that every data point is returned to the set at the sample completion. In this case, the same data point could appear multiple times in the same sample. Another example of bootstrapping can be seen in Barbara (*4*). When we analyze the

mean of the packet sizes we get 523.9. A histogram representing the bootstrapped mean is shown in figure 4.

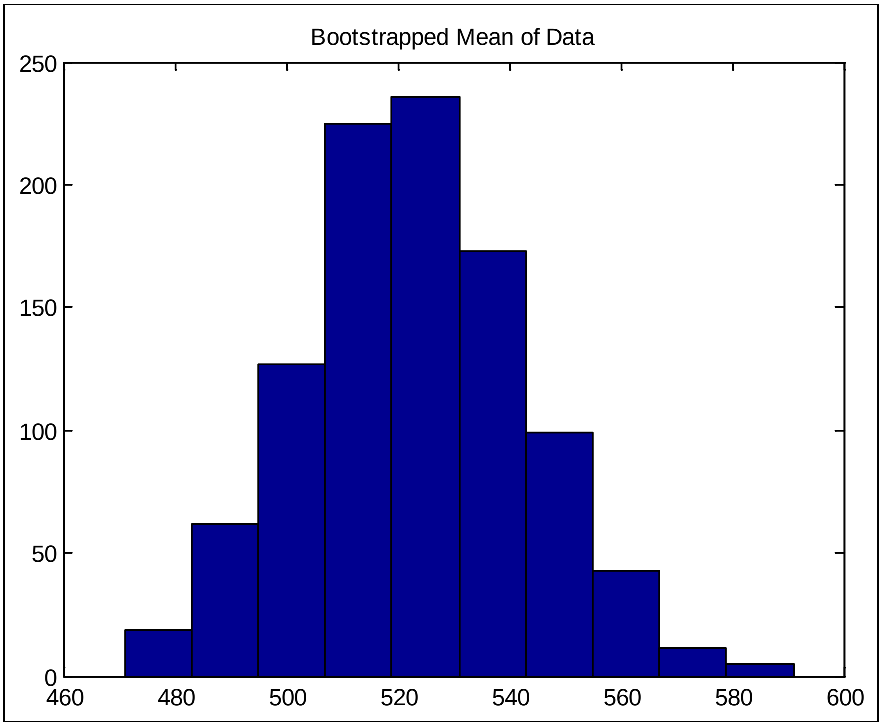

Figure 4. Histogram representing the bootstrapping of the average of packet sizes.

The bootstrap of the mean as seen in the histogram shows the expected variation of the mean. It is easy to see that the most occurring mean is near the actual mean of the data.

Next we made a histogram of the bootstrapped entropy to show the expected entropy of the data set. This is represented in figure 5.

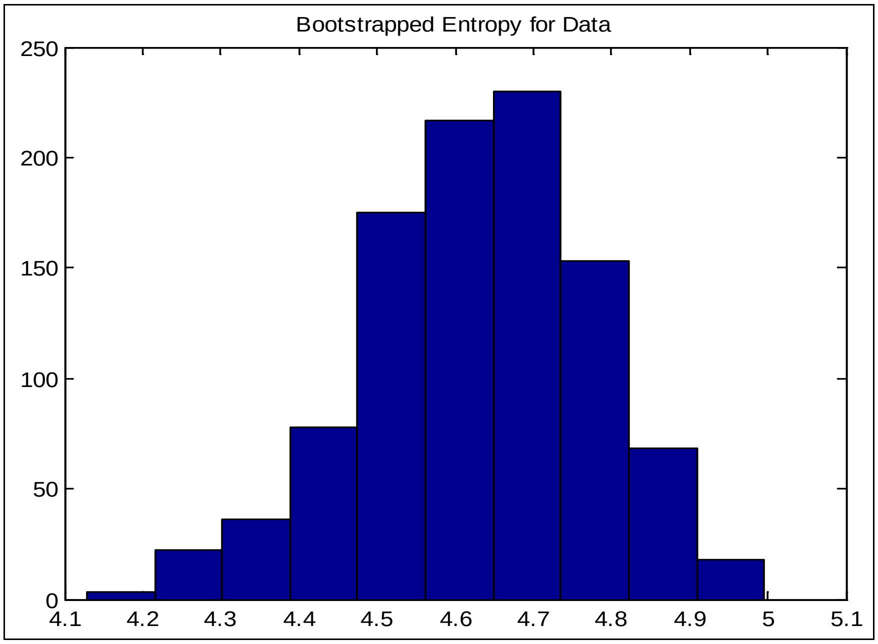

Figure 5. Bootstrapped entropy for data.

The actual entropy of the data (5.05) is not included in the bounds of the bootstrapped entropy. This was most likely caused by the paucity of the data. There were only 82 observations for a range that spanned 800 units. If we were to use these histograms to detect an anomaly, we would look for an entropy value that lay below 4.2 and above 5.00. We would deem entropy values in those ranges as strange behavior. The data collected was relevant to the experiment and helped show how the anomaly detection method could potentially work. Given more data, we would be able to show in a clearer fashion how this approach would effectively detect attacks that affected the expected packet size of SIP packets.

6. Conclusion

Entropy can be a useful tool in the detection of novel attacks. Entropy can be used to detect a change in the basic probability structure of the data. In order to accomplish this, a collection of data points (e.g., packet dump) would be needed. It is evident throughout this report and its findings, that measurements of entropy can identify strange occurrences in a collection of data. Entropy alone is not sufficient; however, to identify what exactly caused the anomaly. Follow-on work could include the use of a statistical model (e.g., Mahalanobis Distance) to identify the individual anomaly.

The use of detecting computer intrusion via entropy has been investigated and is a plausible idea to apply to IDSs. Entropy can detect strange occurrences in a data stream over a specified time frame. One of the many applications of this approach is in detecting data exfiltration, which can

occur at any stage of a VoIP conversation (*5*). Entropy can be applied in various ways to examine data, but it is not a standalone IDS. It offers a theoretical, yet practical approach for the detection of abnormal patterns of behavior.

7. References

1. Esteban M. D.; Morales, D. *A Summary of Entropy Statistics*, Universidad Complutense de Madrid, 28040, Madrid, Spain, 1995.

2. Wikipedia, http://en.wikipedia.org/wiki/Session Initiation Protocol, (accessed June 2009)

3. Boncelet, Charles; Marvel, Lisa M. Steganalysis of Packet Level VoIP Stenography. *Mobile Multimedia/Image Processing, Security, and Applications 2010 Conference, SPIE Defense, Security, and Sensing Symposium*, Orlando FL, April 5 5—9, 2010.

4. Barbara, Daniel; Li Yi; Couto Julia; Lin, Jia-Ling; Sushil Jajodia: Bootstrapping a Data Mining Intrusion Detection System Published by Association for Computing Machinery (ACM). *Proceedings of the 2003 ACM symposium on Applied Computing*, Melbourne, Florida, 2003; pp 421—725.

5. J.C. Pelaez; E.B. Fernandez. VoIP Network Forensic Patterns Published by the U.S. Army Research Laboratory. *Proceedings of the Fourth International Multi-Conference on Computing in the Global Information Technology (ICCGI 2009)*, Cannes, France, August 23—29, 2009.

List of Symbols, Abbreviations, and Acronyms

ARL	U.S. Army Research Laboratory
IDSs	Intrusion Detection Systems
IP	Internet Protocol
SIP	Session Initiation Protocol
UA	user authentication
UPS	Unique Packet Sizes
URL	Uniform Resource Locator
VoIP	Voice over Internet Protocol

NO. OF COPIES	ORGANIZATION
1 ELEC	ADMNSTR DEFNS TECHL INFO CTR DTIC OCP 8725 JOHN J KINGMAN RD STE 0944 FT BELVOIR VA 22060-6218
1	US ARMY RSRCH LAB IMNE ALC HRR 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	US ARMY RSRCH LAB RDRL CIM L 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	US ARMY RSRCH LAB RDRL CIM P 2800 POWDER MILL RD ADELPHI MD 20783-1197
4	US ARMY RSRCH LAB RDRL CI B FORNOFF D DENT F LUDDEN J GOWENS 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	US ARMY RSRCH LAB RDRL CI R NAMBURU 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	US ARMY RSRCH LAB RDRL CII B BROOME 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	US ARMY RSRCH LAB RDRL CII B L TOKARCIK 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	US ARMY RSRCH LAB ATTN RDRL CIN C ARNOLD 2800 POWDER MILL RD ADELPHI MD 20783-1197

NO. OF COPIES	ORGANIZATION
5	US ARMY RSRCH LAB RDRL CIN D A BENCIVENGA D DAVIS K BENSON K LONG W GLODEK 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	US ARMY RSRCH LAB RDRL CIN G RACINE 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	US ARMY RSRCH LAB RDRL CIN S A CLARK 2800 POWDER MILL RD ADELPHI MD 20783-1197
4	US ARMY RSRCH LAB RDRL CIN T B RIVERA N IVANIC R HARDY R PRESSLEY 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	US ARMY RSRCH LAB RDRL CIO S NILES 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	NORTHERN MICHIGAN UNIV ATTN J ANDERTON 3001 NEW SCIENCE FACILITY MARQUETTE MI 49855
1	NORTHERN MICHIGAN UNIV ATTN L ABABNEH 3115 NEW SCIENCE FACILITY MARQUETTE MI 49855
1	NORTHERN MICHIGAN UNIV ATTN R REGIS 3009 NEW SCIENCE FACILITY MARQUETTE MI 49855

NO. OF COPIES	ORGANIZATION
4	NORTHERN MICHIGAN UNIV ATTN NSF 1001A J D PHILLIPS ATTN NSF 1121 G JALILAN ZALMAI ATTN NSF 1131 R APPLETON ATTN NSF 1127 M KOWALCZYK MARQUETTE MI 49855
1	NORTHERN MICHIGAN UNIV ATTN W TIREMAN 2505 WEST SCIENCE BLDG MARQUETTE MI 49855
1	UNIV OF DELAWARE DEPT OF ELECT ENGRG ATTN C BONCELET JR NEWARK DE 19716

<u>ABERDEEN PROVING GROUND</u>

23	DIR USARL RDRL CIM G (BLDG 4600) AMSRD AAR AEF T M ANDRIOLO RDRL CIH C NIETUBICZ RDRL CII C A NEIDERER B BODT J DUMER RDRL CIN D G W THOMPSON (4 COPIES) B RESCHLY C ELLIS J PELAEZ L M MARVEL P GUARINO A PRESSLEY RDRL WML A B FLANDERS D W WEBB A THOMPSON RDRL CIH N C ADAMS RDRL CII C A BORNSTEIN RDRL WML B P KASTE RDRL SLB D J COLLINS